Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2009 *Paper* 50

Validation of Differential Gene Expression Algorithms: Application Comparing Fold Change Estimation to Hypothesis Testing

David R. Bickel* Corey M. Yanofsky†

*University of Ottawa, dbickel@uottawa.ca

†University of Ottawa

Validation of Differential Gene Expression Algorithms: Application Comparing Fold Change Estimation to Hypothesis Testing

David R. Bickel and Corey M. Yanofsky

Abstract

Sustained research on the problem of determining which genes are differentially expressed on the basis of microarray data has yielded a plethora of statistical algorithms, each justified by theory, simulation, or ad hoc validation and yet differing in practical results from equally justified algorithms. The widespread confusion on which method to use in practice has been exacerbated by the finding that simply ranking genes by their fold changes sometimes outperforms popular statistical tests.

Algorithms may be compared by quantifying each method's error in predicting expression ratios, whether such ratios are defined across microarray channels or between two independent groups. For the data sets considered, estimating prediction error by cross validation demonstrates that empirical Bayes methods based on the lognormality assumption tend to outperform both a nonparametric method and algorithms based on selecting genes by their fold changes. The general comparison methodology is applicable to both single-channel and dual-channel microarrays.

As a theoretically sound method of estimating prediction error from observed expression levels, cross validation provides an empirical approach to assessing methods for detecting differential gene expression.

1 Background

Continual invention of new microarray data analysis algorithms for the identification of which genes express differently across two groups calls for objectively comparing the performance of existing algorithms [1]. While there have been thorough empirical comparisons between supervised learning methods of classifying microarrays, comparisons between methods of detecting differential gene expression tend to depend more on theory and simulation than on biological data; for respective examples, see [2] and [3].

By showing that empirical method validation is possible even for algorithms of detecting differential gene expression, a report of the MicroArray Quality Control (MAQC) project [4] may mark a turning point in the methodology of comparing of statistical methods designed to identify differential gene expression on the basis of microarray observations. A notable feature of this "concordance" (percentage of overlapping genes) method is its validation on the basis of the microarray data without resorting to other types of data. Validation by non-microarray information such as RT-PCR measurements of gene expression or public pathway/functional information on genes does have great value in overcoming shortcomings in microarray platforms [5]. For that very reason, however, such validation has markedly less value in judging the performance of statistical methods of detecting differential gene expression. For example, the inability of RT-PCR to validate a microarray prediction of differential gene expression might indicate a problem with the statistical assumptions used to make the prediction, but it may instead reflect a problem with cross hybridization due to the microarray platform. Participants in the MAQC project avoided such confounding between microarray platform effects and statistical method effects by quantifying the degree of overlap between gene lists produced by an algorithm on the basis of two independent data sets [4]. Although a significant step forward, this way of comparing algorithms, like that of [6], requires examining gene lists of given sizes, which is why Chen *et al.* [7] consider the concordance to be too unstable for use as an algorithm performance criterion. Without depending on arbitrarily selected numbers of genes, the platform-algorithm confounding may be overcome by instead using a test set of microarrays to validate predictions made on the basis of a separate training set of microarrays, as explained in Section 2 and illustrated in Section 3; implications are discussed in Section 4.

2 Methods

2.1 Gene selection algorithms

If a gene is known to be differentially expressed at a certain level on average, then that level would predict future measurements of gene expression better than would making such predictions on the assumption that there is on average no differential expression. Likewise, if a gene is known to be equivalently expressed, then using an expression level of 0 or an expression ratio of 1 would predict future measurements better than making such predictions on the assumption that there is some differential expression. Thus, a method of selecting genes as differentially expressed may be judged by estimating its ability to predict future measurements of gene expression. This estimation may be carried out by a process of *cross validation*: the microarrays are divided between a *training set* used to determine which genes the method considers differentially expressed and a *test set* used to estimate how well such results would agree with future measurements.

The strategy of assessing gene selection algorithms by estimated prediction error may be more precisely specified in mathematical notation. Let $x_{i,j}$ denote the logarithm of the measured expression intensity or ratio of intensities of the ith of m genes in the jth of n biological replicates of the control or reference group; each value of $x_{i,j}$ may represent an average over technically replicated microarrays; $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,n})$; $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)^T$. Likewise, $x'_{i,j}$ denotes the logarithm of the measured expression intensity or ratio of intensities of the ith gene in the jth of n' biological replicates of the treatment or perturbation group; $\mathbf{x}'_i = \left(x'_{i,1}, x'_{i,2}, ..., x'_{i,n'}\right)$; $\mathbf{x}' = (\mathbf{x}'_1, \mathbf{x}'_2, ..., \mathbf{x}'_m)^T$. The observations $x_{i,j}$ and $x'_{i,j}$ are realizations of the random variables X_i and X'_i, respectively. The ith gene is called *equivalently expressed* if $\langle X'_i - X_i \rangle = 0$ or *differentially expressed* if $\langle X'_i - X_i \rangle \neq 0$, where the angular brackets denote the expectation value over the sample space. In hypothesis testing parlance, the null hypothesis associated with the ith gene is H_i : $\langle X'_i - X_i \rangle = 0$.

A gene selection algorithm α returns $\pi_\alpha(H_i|\mathbf{x}', \mathbf{x})$, an estimate of the posterior probability that the ith gene is equivalently expressed; it follows that $1 - \pi_\alpha(H_i|\mathbf{x}', \mathbf{x})$ is the algorithm's probability that the gene is differentially expressed across the perturbation and reference groups. Many algorithms [8–17] give $\pi_\alpha(H_i|\mathbf{x}', \mathbf{x})$ directly as a local false discovery rate estimate [18, 19], whereas traditional false discovery rate estimates and other non-Bayesian algorithms in effect assign $\pi_\alpha(H_i|\mathbf{x}', \mathbf{x})$ a value of either 0 or 1, depending on whether or not a gene is considered differentially expressed at a given threshold. For example, the practice of considering a gene differentially expressed if $\exp(|\bar{x}'_i - \bar{x}_i|)$, its estimated *fold change*, is at least ϕ may be expressed as

$$\pi_{\text{foldchange} > \phi}(H_i|\mathbf{x}', \mathbf{x}) = \begin{array}{ll} 0 & \text{if } |\bar{x}'_i - \bar{x}_i| \geq \log(\phi) \\ 1 & \text{if } |\bar{x}'_i - \bar{x}_i| < \log(\phi) \end{array} \tag{1}$$

with $\phi > 0$, $\bar{x}'_i = \sum_{j=1}^{n'} x'_{i,j}/n'$, and $\bar{x}_i = \sum_{j=1}^{n} x_{i,j}/n$. The discontinuity can be removed by introducing smooth functions such as

$$\pi_{\text{fold change shrinkage}}(H_i|\mathbf{x}', \mathbf{x}) = e^{-\left(\exp\left(|\bar{x}'_i - \bar{x}_i|\right) - 1\right)}. \tag{2}$$

The trivial algorithms

$$\pi_{\text{all nulls true}}(H_i|\mathbf{x}', \mathbf{x}) = 1, \tag{3a}$$

$$\pi_{\text{all nulls false}}(H_i|\mathbf{x}', \mathbf{x}) = 0, \tag{3b}$$

which completely ignore the data, will serve as informative points of reference.

Some of the empirical Bayes algorithms implemented in two R packages [20] are considered here [21–23]. From calculations based on a moderated (regularized) t-statistic that are performed by the R package *limma* [21], one may readily obtain $p_i(\tilde{t})$, a one-sided p-value of the ith null hypothesis; $\mathbf{p}(\tilde{t}) = (p_1(\tilde{t}), p_2(\tilde{t}), ..., p_m(\tilde{t}))$. Given the moderated t-statistics and $\pi(H_0)$, the proportion of genes expected to be equivalently expressed, limma also computes $\log\omega_i(\pi(H_0))$, the estimated logarithm of the posterior odds that gene i is differentially expressed rather than equivalently expressed, from which the local false discovery rate may be readily obtained as $(1 + \omega_i(\pi(H_0)))^{-1}$. Since, for use with the log-odds, the author of the algorithm does not recommend computing $\pi(H_0)$ using limma's *convest* function (Gordon Smyth, personal communication, 27 Oct. 2007), we instead iterated the log-odds function until convergence by adapting a method [24] originally proposed for another empirical Bayes algorithm [25]:

1. Let $\pi_1(H_0) = 90\%$ and initialize k to 1.
2. Increment k by 1.
3. Let $\pi_k(H_0) = \sum_{i=1}^{m} (1 + \omega_i(\pi_{k-1}(H_0)))^{-1}/m$.
4. Repeat Steps 2-3 until the absolute value of the proportion difference is sufficiently small, i.e., $|\pi_k(H_0) - \pi_{k-1}(H_0)| < 1/1000$, or until the sign of the proportion difference changes, i.e., $(\pi_k(H_0) - \pi_{k-1}(H_0))(\pi_{k-1}(H_0) - \pi_{k-2}(H_0)) < 0$. The number of iterations performed until such convergence is denoted by K.
5. Let $\pi(H_0) = \pi_K(H_0)$.

Based on that value of $\pi(H_0)$, the estimated probability of equivalent expression is

$$\pi_{\text{moderated t stat. with limma}}(H_i|\mathbf{x}', \mathbf{x}) = \frac{1}{1 + \omega_i(\pi(H_0))}. \tag{4}$$

Also using standard distributions of test statistics under the null hypothesis, the R package *locfdr* [22] maps $\mathbf{p}$, a vector of single-tailed p-values for all genes, to estimates of a local false discovery rate, $\pi_{\text{locfdr}}(H_i, \mathbf{p}|\mathbf{x}', \mathbf{x})$. The use of moderated t-statistics is incorporated by

$$\pi_{\text{moderated t stat. with locfdr}}(H_i|\mathbf{x}', \mathbf{x}) = \pi_{\text{locfdr}}\left(H_i, \mathbf{p}\left(\tilde{t}\right)|\mathbf{x}', \mathbf{x}\right). \tag{5}$$

More commonly, $\mathbf{p}(t)$, a vector of standard (1- or 2-sample) t-test p-values, which also assume the normality of $X_i' - X_i$, or $\mathbf{p}(w)$, a vector of (signed-rank or rank-sum) Wilcoxon test p-values, which do not assume normality, yield local false discovery rate estimates

$$\pi_{\text{t stat. with locfdr}}(H_i|\mathbf{x}', \mathbf{x}) = \pi_{\text{locfdr}}(H_i, \mathbf{p}(t)|\mathbf{x}', \mathbf{x}), \tag{6a}$$

$$\pi_{\text{Wilcoxon stat. with locfdr}}(H_i|\mathbf{x}', \mathbf{x}) = \pi_{\text{locfdr}}(H_i, \mathbf{p}(w)|\mathbf{x}', \mathbf{x}). \tag{6b}$$

Alternatively, the *locfdr* package can employ an empirical maximum likelihood estimate of the null distribution [23] for computation of the local false discovery rate estimate $\pi_{\text{emp. null}}(H_i, \mathbf{p}|\mathbf{x}', \mathbf{x})$:

$$\pi_{\text{t stat. with emp. null}}(H_i|\mathbf{x}', \mathbf{x}) = \pi_{\text{emp. null}}(H_i, \mathbf{p}(t)|\mathbf{x}', \mathbf{x}), \tag{7a}$$

$$\pi_{\text{Wilcoxon stat. with emp. null}}(H_i|\mathbf{x}', \mathbf{x}) = \pi_{\text{emp. null}}(H_i, \mathbf{p}(w)|\mathbf{x}', \mathbf{x}). \tag{7b}$$

Whereas the empirical Bayes methods provide approximations to a posterior probability of a hierarchical Bayesian class of models, we included comparisons to the posterior probability $\pi_{\text{Bayes factor}}(H_i|\mathbf{x}', \mathbf{x})$ under a non-hierarchical set of models. The data densities under the non-hierarchical models are based on the same assumptions as those of standard linear regression: unconstrained data means under the alternative hypothesis (differential expression) and, for each gene, normal IID noise and equal variance within each group in the unpaired case. The posterior odds of differential expression under these models are

$$\omega_{i,\text{ Bayes factor}} = \frac{P(H_i)}{P\left(\widetilde{H}_i\right)} \frac{P(d\mathbf{x}', d\mathbf{x}|H_i)}{P\left(d\mathbf{x}', d\mathbf{x}|\widetilde{H}_i\right)}, \tag{8}$$

where $\widetilde{H}_i$ represents the hypothesis of differential expression and $P(d\mathbf{x}', d\mathbf{x}|h)$ is the prior predictive density or integrated likelihood under hypothesis h. The left-hand side of equation (8) is the posterior odds of equivalent expression to differential expression; on the right-hand side, the first factor is the prior odds of equivalent expression to differential expression, and the second factor is known as the *Bayes factor*. Since we take $P(H_i) = P\left(\widetilde{H}_i\right) = 1/2$, our posterior odds is equal to the Bayes factor; thus putting equal prior mass on each hypothesis does not share the conservatism of the above empirical Bayes methods. The Supplementary Information gives the analytical derivation of the resulting posterior probability, which may be expressed in terms of some additional notation. Define

$$k_1 = \sqrt{n+1}; k_2 = 1; k_3 = \frac{n}{2} \tag{9}$$

$$\xi_i^2 = \frac{n}{n+1}\left(\bar{x}_i' - \bar{x}_i\right)^2; SSR_{H_i} = \sum_{j=1}^{n}\left(x_{i,j}' - x_{i,j}\right)^2;$$

$$SSR_{\widetilde{H}_i} = \sum_{j=1}^{n}\left(\left(x_{i,j}' - x_{i,j}\right) - \left(\bar{x}_i' - \bar{x}_i\right)\right)^2$$

if $n = n'$ and $x_{i,j}'$ is paired with $x_{i,j}$, or

$$k_1 = \sqrt{n+n'+nn'}; k_2 = \sqrt{n+n'}; k_3 = \frac{n+n'-1}{2} \tag{10}$$

$$\xi_i^2 = \frac{nn'}{n+n'+nn'}\left(\bar{x}_i' - \bar{x}_i\right)^2$$

$$SSR_{H_i} = \sum_{j=1}^{n'}\left(x_{i,j}' - \frac{n\bar{x}_i + n'\bar{x}_i'}{n+n'}\right)^2 + \sum_{j=1}^{n}\left(x_{i,j} - \frac{n\bar{x}_i + n'\bar{x}_i'}{n+n'}\right)^2$$

$$SSR_{\widetilde{H}_i} = \sum_{j=1}^{n}\left(x_{i,j}' - \bar{x}_i'\right)^2 + \sum_{j=1}^{n}\left(x_{i,j} - \bar{x}_i'\right)^2$$

if X_i' and X_i are independent. Then the posterior probability is given by

$$\pi_{\text{Bayes factor}}\left(H_i|\mathbf{x}', \mathbf{x}\right) = \frac{1}{1+\omega_{i,\text{ Bayes factor}}}, \tag{11}$$

$$\omega_{i,\text{Bayes factor}} = \frac{k_1\left(\xi_i^2 + SSR_{\widetilde{H}_i}\right)^{k_3}}{k_2\left(SSR_{H_i}\right)^{k_3}}. \tag{12}$$

We also applied two "information criteria" used in model selection to estimate the posterior probability; the information criteria were applied to the same linear regression framework used in the above Bayes factor computation. In model selection terminology, each criterion selects either model H_i or model $\widetilde{H}_i$ for the ith gene, but we instead averaged the estimates corresponding to the two models for each gene as follows. We first applied the Bayesian Information Criterion (BIC) [26]. Up to a factor of $-1/2$ and a constant term, the BIC approximates the logarithm of the prior predictive probability density given a statistical model and a sufficiently diffuse proper prior distribution under the given model without requiring specification of such a prior. With a prior mass on each model considered, the BIC leads to an approximation of a posterior probability that is less conservative than that of the above Bayes factor. For paired data, the BIC for each hypothesis is

$$\text{BIC}\left(\widetilde{H}_i\right) = n\log\left(\frac{SSR_{\widetilde{H}_i}}{n}\right) + 2\log n, \tag{13}$$

$$\text{BIC}\left(H_i\right) = n\log\left(\frac{SSR_{H_i}}{n}\right) + \log n, \tag{14}$$

with SSR_h as defined in (9); for independent data, the BIC values are

$$\text{BIC}\left(\widetilde{H}_i\right) = (n+n')\log\left(\frac{SSR_{\widetilde{H}_i}}{n+n'}\right) + 3\log\left(n+n'\right), \tag{15}$$

$$\text{BIC}\left(H_i\right) = (n+n')\log\left(\frac{SSR_{H_i}}{n+n'}\right) + 2\log\left(n+n'\right), \tag{16}$$

with SSR_h as defined in (10). Since we again use $P\left(H_i\right) = P\left(\widetilde{H}_i\right)$, the BIC approximation of the posterior odds ($\omega_{i,\text{BIC}}$) is equal to its approximation of Bayes factors corresponding to a wide class of priors on the model parameters. Transformed from the logarithmic scale to the probability scale [27], the result is an

equation of the same form as (11),

$$\pi_{\mathrm{BIC}}\left(H_i|\mathbf{x}',\mathbf{x}\right) = \frac{1}{1+\omega_{i,\ \mathrm{BIC}}}, \tag{17}$$

$$\omega_{i,\mathrm{BIC}} = \frac{\exp\left[-\frac{1}{2}\mathrm{BIC}\left(\widetilde{H}_i\right)\right]}{\exp\left[-\frac{1}{2}\mathrm{BIC}\left(H_i\right)\right]}. \tag{18}$$

The second information criterion we assessed was the Akaike Information Criterion corrected for small samples ($\mathrm{AIC_c}$). While $-\mathrm{AIC_c}/2$ plus a constant term is in general only an approximately unbiased estimator of the expected Kullback-Leibler distance between the model/hypothesis and the unknown true data generating distribution [28], it is exactly unbiased for linear regression models with normal errors [29], a class that includes the present non-hierarchical models. Under the name of *Akaike weights*, it and other AIC-like criteria have been used to generate predictions that take model uncertainty into account in a manner exactly analogous to Bayesian model averaging [28], giving rise to an equation of the same form as (17). For paired data, the $\mathrm{AIC_c}$ values of the hypotheses or models are

$$\mathrm{AIC_c}\left(\widetilde{H}_i\right) = n\log\left(\frac{SSR_{\widetilde{H}_i}}{n}\right)+\frac{4n}{n-3}, \tag{19}$$

$$\mathrm{AIC_c}\left(H_i\right) = n\log\left(\frac{SSR_{H_i}}{n}\right)+\frac{2n}{n-2}, \tag{20}$$

with SSR_h as defined in (9); for independent data, the $\mathrm{AIC_c}$ values are

$$\mathrm{AIC_c}\left(\widetilde{H}_i\right) = (n+n')\log\left(\frac{SSR_{\widetilde{H}_i}}{n+n'}\right)+\frac{6\,(n+n')}{n+n'-4}, \tag{21}$$

$$\mathrm{AIC_c}\left(H_i\right) = (n+n')\log\left(\frac{SSR_{H_i}}{n+n'}\right)+\frac{4\,(n+n')}{n+n'-3}, \tag{22}$$

with SSR_h as defined in (10). Transforming from the logarithmic scale yields the effective probability

$$\pi_{\mathrm{AIC_c}}\left(H_i|\mathbf{x}',\mathbf{x}\right) = \frac{1}{1+\omega_{i,\mathrm{AIC_c}}}, \tag{23}$$

where

$$\omega_{i,\mathrm{AIC_c}} = \frac{\exp\left[-\frac{1}{2}\mathrm{AIC_c}\left(\widetilde{H}_i\right)\right]}{\exp\left[-\frac{1}{2}\mathrm{AIC_c}\left(H_i\right)\right]}$$

is the ratio of Akaike weights.

2.2 Methods of assessing gene selection algorithms

Algorithm α's best prediction of future values of $X_i' - X_i$ is $\pi_\alpha\left(H_i|\mathbf{x}',\mathbf{x}\right)(0) + \left(1-\pi_\alpha\left(H_i|\mathbf{x}',\mathbf{x}\right)\right)\left(\bar{x}_i' - \bar{x}_i\right)$; this approximation of posterior mean degree of expression has been compared to a method of correcting estimates for gene selection bias [30]. The corresponding estimate of the prediction error is

$$\hat{\epsilon}_{\alpha,i} = \frac{1}{n}\sum_{j=1}^{n}\left[\begin{array}{c}\left(x'_{i,j}-x_{i,j}\right)-\\ \left(1-\pi_\alpha\left(H_i|\mathbf{x}'_{(-j)},\mathbf{x}_{(-j)}\right)\right)\left(\bar{x}'_{i,(-j)}-\bar{x}_{i,(-j)}\right)\end{array}\right]^2 \tag{24}$$

if $n = n'$ and $x'_{i,j}$ is paired with $x_{i,j}$ or

$$\hat{\epsilon}_{\alpha,i} = \frac{1}{nn'}\sum_{j,j'=1}^{n,n'}\left[\begin{array}{c}\left(x'_{i,j'}-x_{i,j}\right)-\\ \left(1-\pi_\alpha\left(H_i|\mathbf{x}'_{(-j')},\mathbf{x}_{(-j)}\right)\right)\left(\bar{x}'_{i,(-j')}-\bar{x}_{i,(-j)}\right)\end{array}\right]^2 \tag{25}$$

if X'_i and X_i are independent, where $(-j)$ means the jth replicate is omitted:

$$\mathbf{x}'_{(-j)} = \begin{pmatrix} x'_{1,1} & \cdots & x'_{1,j-1} & x'_{1,j+1} & \cdots & x'_{1,n'} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_{m,1} & \cdots & x'_{m,j-1} & x'_{m,j+1} & \cdots & x'_{m,n'} \end{pmatrix}, \tag{26a}$$

$$\mathbf{x}_{(-j)} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,j-1} & x_{1,j+1} & \cdots & x_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & \cdots & x_{m,j-1} & x_{m,j+1} & \cdots & x_{m,n} \end{pmatrix}, \tag{26b}$$

$\bar{x}'_{i,(-j)} = \left(\sum_{J=1}^{n'} x'_{i,j} - x'_{i,j}\right) / (n'-1)$, and $\bar{x}_{i,(-j)} = \left(\sum_{J=1}^{n} x_{i,j} - x_{i,j}\right) / (n-1)$. The error relative to always predicting that $X'_i - X_i = 0$ is

$$\hat{\varepsilon}_{\alpha,i} = \frac{\hat{\epsilon}_{\alpha,i}}{\hat{\epsilon}_{\text{all nulls true},i}}. \tag{27}$$

Three ways to average the estimated prediction error of an algorithm over all genes are

$$\circ_\alpha = (\text{relative error mode})_\alpha = \text{HSM}\left(\hat{\varepsilon}_{\alpha,1}, \hat{\varepsilon}_{\alpha,2}, ..., \hat{\varepsilon}_{\alpha,m}\right), \tag{28}$$

$$\triangle_\alpha = (\text{relative error mean})_\alpha = \frac{1}{m}\sum_{i=1}^{m} \hat{\varepsilon}_{\alpha,i}, \tag{29}$$

$$+_\alpha = (\text{absolute error})_\alpha = \frac{\sum_{i=1}^{m} \hat{\epsilon}_{\alpha,i}}{\sum_{i=1}^{m} \hat{\epsilon}_{\text{all nulls true},i}}; \tag{30}$$

the half-sample mode (HSM) is the estimator of the mode studied in [31] and implemented as the *hsm* function in the *modeest* package of R.

Jeffery *et al.* [32] also used a cross-validation approach to estimate the predictive error of a variety of gene selection algorithms, but with microarray classification error rather than equations (28) as the performance criterion. Such classification error depends not only on the gene selection algorithm, but also on the particular classifier for which that algorithm selects features. Since our interest lies strictly in identifying differentially expressed genes, our methods instead quantify performance in terms of predicting new measurements.

3 Results

To illustrate the proposed methods of quantifying the performance of gene selection algorithms, we applied them to two example data sets, one relevant to agriculture and the other to medicine.

3.1 Agricultural data

Dual-channel microarrays were used to measure in tomatoes the expression ratios (mutant/wild type) of $m = 13,440$ genes at the breaker stage of ripening and at 3 and 10 days thereafter [33]. Each of the later two stages has six biological replicates ($n = 6$), but one of the biological replicates is missing at the breaker stage of ripening ($n = 5$). The next sub$\Lambda > 1$ compares algorithms of determining which genes are differentially expressed between mutant and wild type at each point in time, whereas Subsection 3.2 uses the same data to instead compare algorithms of determining which genes are differentially expressed between one point in time and another point in time.

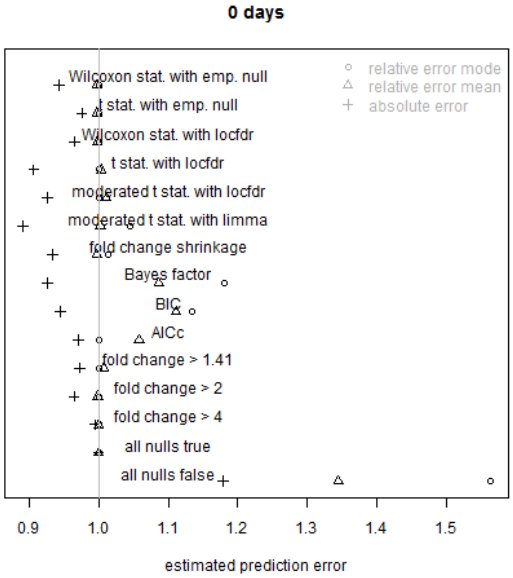

Figure 1: Estimated prediction errors, defined by equations (28), at the breaker stage of ripening. The values of α displayed correspond to the gene selection algorithms of equations (1)-(7).

3.1.1 Pairing across microarray channels

In order to determine the genes for which expected values of logarithms of mutant-to-wild-type ratios differ from 0, let $x'_{i,j}$ be the expression level of the mutant sample with mRNA hybridized to the same microarray as that of a wild type sample with expression level $x_{i,j}$ at 0, 3, or 10 days after the breaker stage. Then $x'_{i,j} - x_{i,j}$ is the logarithm of the observed ratio for the ith gene and jth microarray. Due to this dependence structure, paired (1-sample) t-tests and Wilcoxon signed-rank tests were used to obtain p-values, and equations (28) were used to estimate prediction error. The estimated prediction errors for all algorithms mentioned above are displayed as Figs. 1-3.

3.1.2 Two independent groups

In order to determine which genes differ in mutant-to-wild-type ratios between different periods of time after the breaker stage, let $x'_{i,j'}$ and $x_{i,j}$ the logarithms of ratios observed at two different points in time for gene i and for microarrays j' and j. Since the measurement errors of observations made at one time point are independent of those made at the other time point, 2-sample t-tests and Wilcoxon rank-sum tests were used to obtain p-values, and equations (28) were used to estimate prediction error (Figs. 4-6).

3.2 Biomedical data

MAQC researchers [4] measured gene expression responses to a rat liver treatment on four different platforms: Applied Biosystems, Affymetrix, Agilent, and GE Healthcare. Each data set has six treatment biological replicates and six control biological replicates. As in Subsection 3.1.2, observations in the treatment group are not paired with those of the control group. The Applied Biosystems data set ($m = 26,857$ genes) and Agilent data set ($m = 41,070$ genes) were used to assess gene selection criteria on the basis of prediction error (Figs. 7-8). The *limma* method was not applied to the Agilent data set because it contains repeated minimum measurements, which prevent the software from estimating the prior variance.

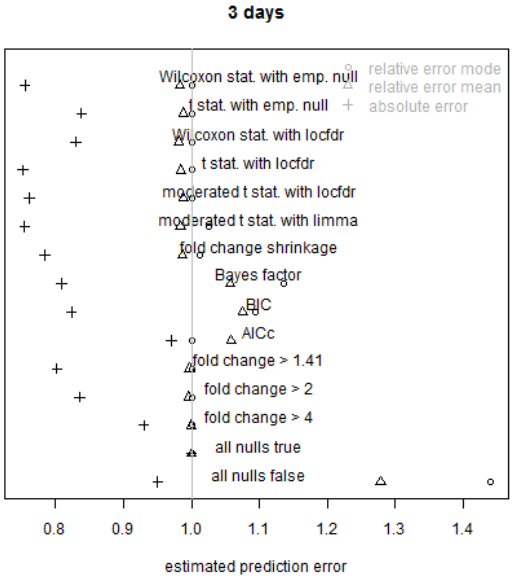

Figure 2: Estimated prediction errors 3 days after the breaker stage of ripening. Error and algorithm definitions are the same as those of Fig. 1.

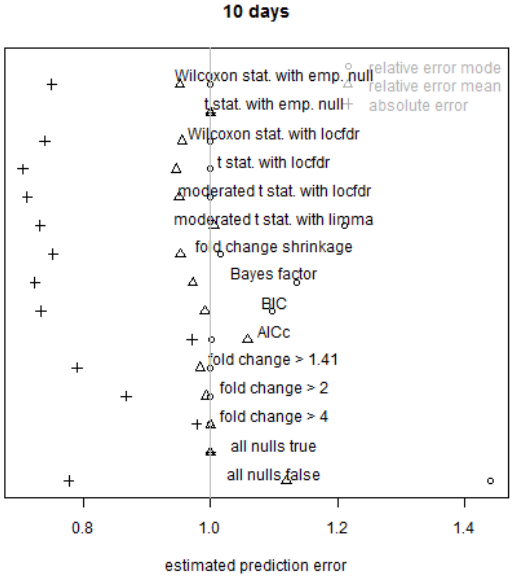

Figure 3: Estimated prediction errors 10 days after the breaker stage of ripening. Error and algorithm definitions are the same as those of Fig. 1.

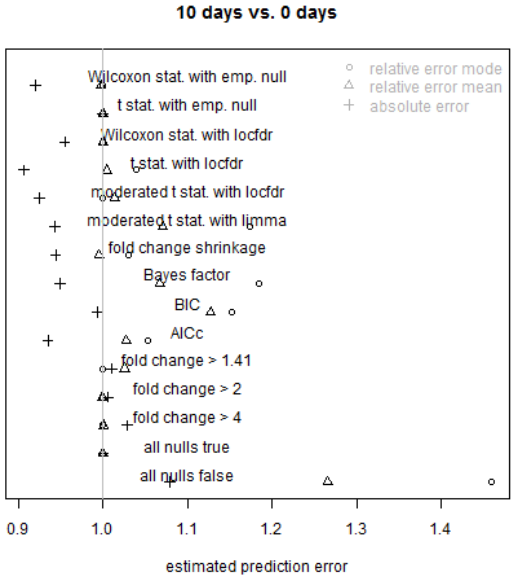

Figure 4: Estimated prediction errors for comparing expression at 10 days to expression at 0 days after the breaker stage of ripening. Error and algorithm definitions are the same as those of Fig. 1.

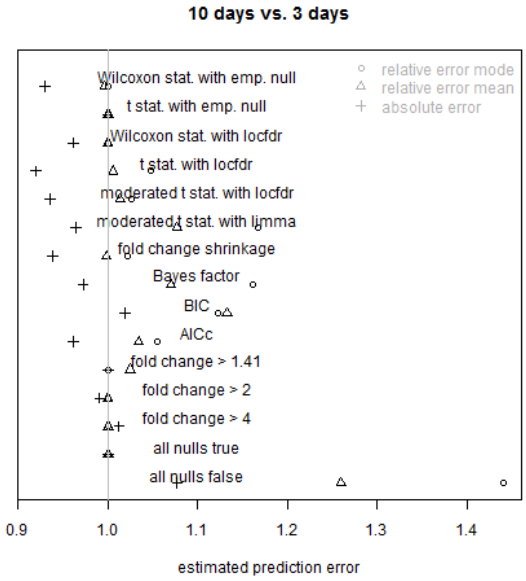

Figure 5: Estimated prediction errors for comparing expression at 10 days to expression at 3 days after the breaker stage of ripening. Error and algorithm definitions are the same as those of Fig. 1.

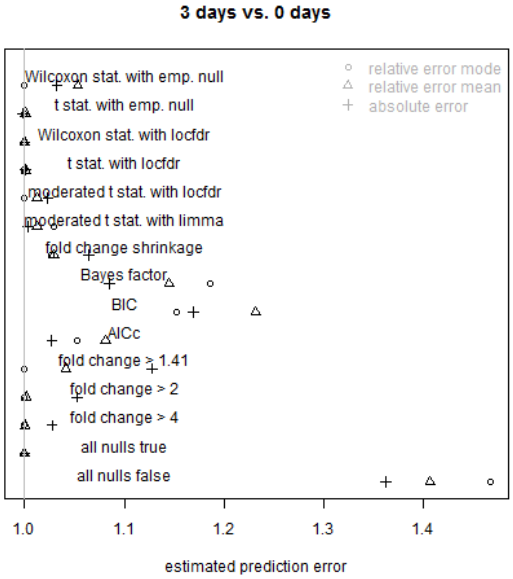

Figure 6: Estimated prediction errors for comparing expression at 3 days to expression at 0 days after the breaker stage of ripening. Error and algorithm definitions are the same as those of Fig. 1.

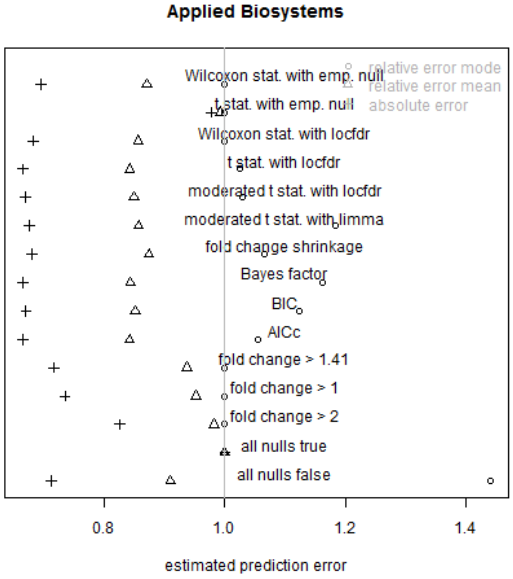

Figure 7: Estimated prediction errors for the Applied Biosystems data set of the rat toxicogenomics study. Error and algorithm definitions are the same as those of Fig. 1.

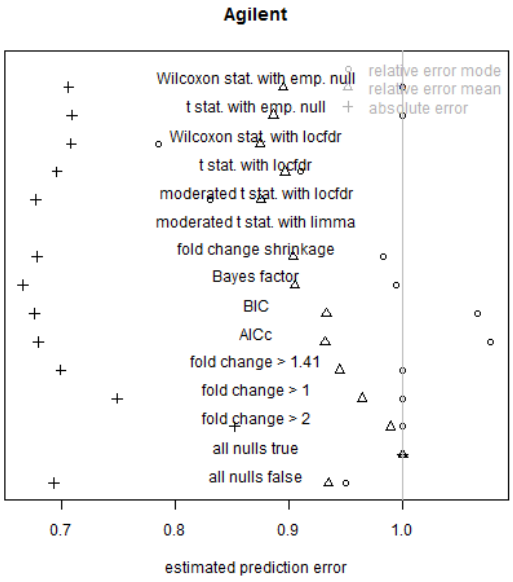

Figure 8: Estimated prediction errors for the Agilent data set of the rat toxicogenomics study. Error and algorithm definitions are the same as those of Fig. 1.

4 Discussion and Conclusions

Some observations made on the basis of the example data sets invite further study. First, the t-statistic methods often had lower estimated prediction errors than the Wilcoxon algorithms, indicating that the distributions of expression ratios are sufficiently close to the lognormal family for parametric testing with small sample sizes. Second, comparing $+_{\text{t stat. with locfdr}}$ to $+_{\text{moderated t stat. with locfdr}}$ suggests that regularization of the t-statistic does not necessarily tend to improve predictive performance. Third, the model selection methods not adjusted for multiple testing (Bayes factor, BIC, and AIC_c) performed well in the experiment designed to have large changes between conditions (Figs. 7-8). Fourth, the hypothesis testing methods generally performed better than methods based on fold-change estimates alone. Even so, fold-change shrinkage (2) performed remarkably well, except in the case in which no genes appear differentially expressed (Fig. 6).

Related to the fourth observation, investigators have been reporting that a heuristic combination of statistical testing and fold-change estimation performs better than does either type of algorithm alone [4,34]. The inferior performance of statistical methods that do not make use of fold-change estimates has been explained both in terms of the high variability in p-values expected with small samples [34] and in terms of a distinction between statistical and biological significance [35]. The latter explanation would call for the incorporation of the lowest fold change considered biologically relevant into the statistical hypotheses under consideration. Recent statistical methods designed to find genes expressed at biologically important levels include those utilizing false discovery rates [36, 37], Bayesian analyses [38, 39], and the likelihood paradigm of measuring the strength of statistical evidence [40]. Thus, researchers need not choose between statistical rigor and incorporation of information about fold change.

Although the proposed cross-validation methodology may be used with data sets with as few as three biological replicates, the variance in cross-validation estimates of the prediction error might be prohibitively high for extremely small samples. For this reason, model-based methods of estimating the prediction error such as parametric posterior predictive inference and parametric bootstrapping [41] also merit attention.

Authors' contributions

DB conceived the study, selected the data sets, applied the fold change and empirical Bayes algorithms, and drafted the manuscript. CY selected the Bayes factor algorithm, implemented the Bayes factor, BIC, and AIC_c algorithms, and helped draft the manuscript.

Acknowledgements

We thank Pei-Chun Hsieh for preparing the biomedical data for analysis and Xuemei Tang for providing the fruit development microarray data. The *Biobase* package of Bioconductor [42] facilitated management of the expression data. This work was partially supported by the Canada Foundation for Innovation (CFI16604), the Ministry of Research and Innovation of Ontario (MRI16604), Dell Computing, and the Faculty of Medicine of the University of Ottawa.

References

1. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: From disarray to consolidation and consensus**. *Nature Reviews Genetics* 2006, **7**:55–65.
2. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *Journal of the American Statistical Association* 2002, **97**(457):77–86.
3. Chen J, van der Laan MJ, Smith MT, Hubbard AE: **A comparison of methods to control Type I errors in microarray studies**. *Statistical Applications in Genetics and Molecular Biology* 2007, **6**:28.
4. Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM, Hurban P, Phillips KL, Xu J, Deng X, Sun YA, Tong W, Dragan YP, Shi L: **Rat toxicogenomic study reveals analytical consistency across microarray platforms**. *Nat Biotech* 2006, **24**(9):1162–1169.
5. Rockett JC, Hellmann GM: **Confirming microarray data - Is it really necessary?** *Genomics* 2004, **83**(4):541–549.
6. Pepe MS, Longton G, Anderson GL, Schummer M: **Selecting differentially expressed genes from microarray experiments**. *Biometrics* 2003, **59**:133–142.
7. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA: **Reproducibility of microarray data: A further analysis of microarray quality control (MAQC) data**. *BMC Bioinformatics* 2007, **8**.
8. Aubert J, Bar-Hen A, Daudin JJ, Robin S: **Determination of the differentially expressed genes in microarray experiments using local FDR**. *BMC Bioinformatics* 2004, **5**.
9. Aubert J, Bar-Hen A, Daudin JJ, Robin S: **Correction: Determination of the differentially expressed genes in microarray experiments using local FDR (BMC Bioinformatics)**. *BMC Bioinformatics* 2005, **6**.
10. Jones LBT, Bean R, McLachlan GJ, Zhu JXI: **Mixture models for detecting differentially expressed genes in microarrays**. *International journal of neural systems* 2006, **16**(5):353–362.
11. Liao JG, Lin Y, Selvanayagam ZE, Shih WJ: **A mixture model for estimating the local false discovery rate in DNA microarray analysis**. *Bioinformatics* 2004, **20**(16):2694–2701.
12. McLachlan GJ, Bean RW, Jones LBT, Zhu JX: **Using mixture models to detect differentially expressed genes**. *Australian Journal of Experimental Agriculture* 2005, **45**(7-8):859–866.
13. Pawitan Y: *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford: Clarendon Press 2001.
14. Ploner A, Calza S, Gusnanto A, Pawitan Y: **Multidimensional local false discovery rate for microarray studies**. *Bioinformatics* 2006, **22**(5):556–565.
15. Pounds S, Morris SW: **Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values**. *Bioinformatics* 2003, **19**(10):1236–1242.
16. Scheid S, Spang R: **A stochastic downhill search algorithm for estimating the local false discovery rate**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2004, **1**(3):98–108.

17. Scheid S, Spang R: **Twilight; a Bioconductor package for estimating the local false discovery rate**. *Bioinformatics* 2005, **21**(12):2921–2922.

18. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes Analysis of a Microarray Experiment**. *J. Am. Stat. Assoc.* 2001, **96**(456):1151–1160.

19. Genovese C, Wasserman L: *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, June 2-6, 2002*, Oxford: Clarendon Press 2002 chap. Bayesian and frequentist multiple testing.

20. R Development Core Team: **R: A Language and Environment for Statistical Computing** 2007.

21. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Statistical Applications in Genetics and Molecular Biology* 2004, **3**.

22. Efron B: **Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis**. *Journal of the American Statistical Association* 2004, **99**(465):96–104.

23. Efron B: **Size, power and false discovery rates**. *Annals of Statistics* 2007, **35**:1351–1351–1377, [http://arxiv.org/abs/0710.2245].

24. Bickel DR: **HighProbability determines which alternative hypotheses are sufficiently probable: Genomic applications include detection of differential gene expression**. *arXiv* 2004, **q-bio/0402049**.

25. Bickel DR: **Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression?** *Statistical Applications in Genetics and Molecular Biology* 2004, **3**:8.

26. Schwarz G: **Estimating the Dimension of a Model**. *The Annals of Statistics* 1978, **6**(2):461–464, [http://www.jstor.org/stable/2958889].

27. Efron B, Gous A, Kass RE, Datta GS, Lahiri P: **Scales of Evidence for Model Selection: Fisher versus Jeffreys**. *Lecture Notes-Monograph Series* 2001, **38**(, Model Selection):208–256, [http://www.jstor.org/stable/4356166].

28. Burnham KP, Anderson D: *Model Selection and Multi-Model Inference*. New York, NY: Springer 2002, [http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20%\&path=ASIN/0387953647].

29. Hurvich CM, Tsai CL: **Regression and Time Series Model Selection in Small Samples**. *Biometrika* 1989, **76**(2):297–307, [http://www.jstor.org/stable/2336663].

30. Bickel DR: **Correcting the estimated level of differential expression for gene selection bias: Application to a microarray study**. *Statistical Applications in Genetics and Molecular Biology* 2008, **7**:10.

31. Bickel DR, Frhwirth R: **On a fast, robust estimator of the mode: comparisons to other robust estimators with applications**. *Computational Statistics and Data Analysis* 2006, **50**:3500–3530.

32. Jeffery IB, Higgins DG, Culhane AC: **Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data**. *BMC Bioinformatics* 2006, **7**.

33. Alba R, Payton P, Fei Z, McQuinn R, Debbie P, Martin GB, Tanksley SD, Giovannoni JJ: **Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development**. *Plant Cell* 2005, **17**(11):2954–2965.

34. Shi L, Jones WD, Jensen RV, Harris SC, Perkins RG, Goodsaid FM, Guo L, Croner LJ, Boysen C, Fang H, Qian F, Amur S, Bao W, Barbacioru CC, Bertholet V, Cao XM, Chu TM, Collins PJ, Fan XH, Frueh FW, Fuscoe JC, Guo X, Han J, Herman D, Hong H, Kawasaki ES, Li QZ, Luo Y, Ma Y, Mei N, Peterson RL, Puri RK, Shippy R, Su Z, Sun YA, Sun H, Thorn B, Turpaz Y, Wang C, Wang SJ, Warrington JA, Willey JC, Wu J, Xie Q, Zhang L, Zhang L, Zhong S, Wolfinger RD, Tong W: **The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies**. *BMC Bioinformatics* 2008, **9**(SUPPL. 9).

35. Chen JJ, Wang SJ, Tsai CA, Lin CJ: **Selection of differentially expressed genes in microarray data analysis**. *Pharmacogenomics Journal* 2007, **7**(3):212–220.

36. Bickel DR: **Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes**. *Bioinformatics (Oxford, England)* 2004, **20**:682–688.

37. Van De Wiel MA, Kim KI: **Estimating the false discovery rate using nonparametric deconvolution**. *Biometrics* 2007, **63**(3):806–815.

38. Lewin A, Richardson S, Marshall C, Glazier A, Aitman T: **Bayesian modeling of differential gene expression**. *Biometrics* 2006, **62**:1–9.

39. Bochkina N, Richardson S: **Tail posterior probability for inference in pairwise and multiclass gene expression data**. *Biometrics* 2007, **63**(4):1117–1125.

40. Bickel DR: **The strength of statistical evidence for composite hypotheses with an application to multiple comparisons**. *unpublished paper, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 49, available at tinyurl.com/7yaysp* 2008.

41. Efron B: **The estimation of prediction error: Covariance penalties and cross-validation**. *Journal of the American Statistical Association* 2004, **99**(467):619–632.

42. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**:R80, [http://genomebiology.com/2004/5/10/R80].

Additional Files

Additional file 1 — Bayes_factor_derivation.pdf

This file contains a heuristic overview and detailed derivation of our Bayes factor approach to calculating probabilities of differential expression.

Supplementary Information

Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing

David R. Bickel, Corey M. Yanofsky

Heuristic overview

This document contains an explicit derivation of the Bayes factor used in the main paper for both paired and unpaired data. In each case, there are two models for the data: the null model in which the gene is equivalently expressed in the two conditions, and the alternative model in which the gene is differentially expressed.

The derivation of the Bayes factor requires two components per model. The first component is the probability distribution of the data conditional on some statistical parameters; this is termed the *likelihood function*. The differential expression model will always have one extra parameter to take into account the fact that the gene's expression level is different across conditions.

The data are always modeled as:

observed datum = *average data level* + *measurement error*.

Here, the average data level is an unknown parameter. Throughout, the measurement errors are assumed to be independent and identically distributed as Gaussian random variables. That is, for all j,

$$p(\varepsilon_j|\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}{\varepsilon_j}^2\right],$$

where ε_j is the measurement error of the j^{th} observation and σ^2 is the data variance .

The second component is the prior distribution of the model parameters, namely, the baseline expression level of the gene and the experimental variability of the data. The prior distribution summarizes everything that is known about the model parameters prior to observing the data. Since the basic expression level of the gene and the variability of the data are unknown, we use standard default priors for them.

The extra parameter in the alternative model measures the amount of differential expression. Here, we use what has been called a *unit-information* prior distribution, that is, a prior distribution that contains exactly as much information as one extra data point. The unit-information prior is weakly informative, so it will not unduly influence the results in favor of either model.

To calculate the Bayes factor, we *marginalize* the model parameters; that is, we integrate the likelihood function with respect to the prior distribution, resulting in a *prior predictive distribution*. The marginalization removes the nuisance parameters from the expression. The Bayes factor is the ratio of the prior predictive distributions under the null and alternative models.

Derivations

Bayes factor for paired data

Suppose $n = n'$ and $x'_{i,j}$ is paired with $x_{i,j}$. Let

$$y_j = x'_{i,j} - x_{i,j}. \qquad \text{(S1)}$$

The hypothesis of equivalent expression is

M_0: $y_j = \varepsilon_j$,

and the hypothesis of differential expression is

M_1: $y_j = \alpha + \varepsilon_j$.

Prior distributions

For both models, we set

$$p(\sigma^2) \propto \frac{1}{\sigma^2}.$$

For M_1, we use the unit information prior

$$p(\alpha|\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\alpha - \mu_\alpha)^2\right].$$

In the main text of the paper, the prior mean μ_α is set to zero.

Null model prior predictive distribution

The prior predictive distribution of the data under M_0 is

$$p(y|M_0) = \int_0^\infty p(\sigma^2) \prod_j p\left(y_j|\mu = 0, \sigma^2\right) d\sigma^2,$$

$$p(y|M_0) = (2\pi)^{-n/2} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left(-\frac{n\overline{y^2}}{2\sigma^2}\right) d\sigma^2,$$

$$p(y|M_0) = (2\pi)^{-n/2} \Gamma\left(\frac{n}{2}\right) \left(n\overline{y^2}\right)^{-n/2}.$$

Alternative model prior predictive distribution

We define in advance:

$$\hat{\alpha} = \frac{(n\bar{y} + \mu_\alpha)}{n+1},$$

$$SSR_1 = (\mu_\alpha - \hat{\alpha})^2 + \sum_j \left(y_j - \hat{\alpha}\right)^2.$$

After some algebra, we can derive

$$SSR_1 = n\left(\overline{y^2} - (\bar{y})^2\right) + \frac{n}{n+1}(\bar{y} - \mu_\alpha)^2.$$

Here, SSR_1 is the effective sum of squares of the residuals under M_1. It is the sum of the *SSR* using the maximum likelihood estimator $\alpha_{MLE} = \bar{y}$ and a term that penalizes disagreement between the MLE and the prior mean.

The prior predictive distribution of the data under M_1 is

$$p(y|M_1) = \int_0^\infty \int_{-\infty}^\infty p(\sigma^2)p(\alpha|\sigma^2)\prod_j p\left(y_j|\mu = \alpha, \sigma^2\right) d\alpha\, d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-(n+1)/2} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{(n+1)}{2}+1} \left(\int_{-\infty}^\infty \exp\left\{ -\frac{1}{2\sigma^2}\left[(\alpha - \mu_\alpha)^2 + \sum_j \left(y_j - \alpha\right)^2 \right] \right\} d\alpha \right) d\sigma^2.$$

Isolating the part of the integrand that is a quadratic expression in α, we complete the square:

$$(\alpha - \mu_\alpha)^2 + \sum_j \left(y_j - \alpha\right)^2$$

$$= \alpha^2 - 2\alpha\mu_\alpha + {\mu_\alpha}^2 + n\alpha^2 - 2\alpha n\bar{y} + n\overline{y^2}$$

$$= (n+1)\alpha^2 - 2\alpha(n\bar{y} + \mu_\alpha) + \left(n\overline{y^2} + {\mu_\alpha}^2\right)$$

$$= (n+1)(\alpha - \hat{\alpha})^2 + \left(n\overline{y^2} - n(\bar{y})^2\right) + {\mu_\alpha}^2 + n(\bar{y})^2 - \frac{n^2(\bar{y})^2 + 2n\bar{y}\mu_\alpha + {\mu_\alpha}^2}{n+1}$$

$$= (n+1)(\alpha - \hat{\alpha})^2 + n\left(\overline{y^2} - (\bar{y})^2\right) + \frac{n{\mu_\alpha}^2 + n^2(\bar{y})^2 + {\mu_\alpha}^2 + n(\bar{y})^2}{n+1} - \frac{n^2(\bar{y})^2 + 2n\bar{y}\mu_{\acute{a}} + {\mu_\alpha}^2}{n+1}$$

$$= (n+1)(\alpha - \hat{\alpha})^2 + n\left(\overline{y^2} - (\bar{y})^2\right) + \frac{1}{n+1}(n(\bar{y})^2 - 2n\bar{y}\mu_\alpha + n{\mu_\alpha}^2)$$

$$= (n+1)(\alpha - \hat{\alpha})^2 + n\left(\overline{y^2} - (\bar{y})^2\right) + \frac{n}{n+1}(\bar{y} - \mu_\alpha)^2$$

$$= (n+1)(\alpha - \hat{\alpha})^2 + SSR_1.$$

Substituting back into the integral, we have

$$p(y|M_1) = (2\pi)^{-(n+1)/2} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{(n+1)}{2}+1} \exp\left(-\frac{SSR_1}{2\sigma^2}\right)\left(\int_{-\infty}^\infty \exp\left[-\frac{n+1}{2\sigma^2}(\alpha - \hat{\alpha})^2\right] d\alpha\right) d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-n/2}(n+1)^{-1/2}\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left(-\frac{SSR_1}{2\sigma^2}\right) d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-n/2}\Gamma\left(\frac{n}{2}\right)(n+1)^{-1/2}(SSR_1)^{-n/2}.$$

The Bayes factor is

$$BF = \frac{p(y|M_0)}{p(y|M_1)} = \sqrt{n+1}\left(\frac{SSR_1}{n\overline{y^2}}\right)^{\frac{n}{2}}. \qquad \text{(S2)}$$

Equations (S1) and (S2) together are equivalent to equations (9), (11), and (12) of the main paper.

Bayes factor for two-sample data

Suppose that $X'_{i,j}$ and $X_{i,j}$ are independent. Define

$$y_j = x_{i,j},\ j = 1, \dots, n, \qquad \text{(S3)}$$

$$y_{n+j} = x'_{i,j},\ j = 1, \dots, n'. \qquad \text{(S4)}$$

The hypothesis of equivalent expression is

M_0: $y_j = \beta + \varepsilon_j, j = 1, \dots, n+n'$,

and the hypothesis of differential expression is

M_1: $y_j = \beta + \varepsilon_j, j = 1, \dots, n,$

$y_j = \alpha + \varepsilon_j, j = n+1, \dots, n+m.$

Preliminaries

To fix notation, let

$$\overline{y^2} = \frac{1}{n+m}\sum_{j=1}^{n+m} {y_j}^2,$$

$$\bar{y} = \frac{1}{n+m}\sum_{j=1}^{n+m} y_j,$$

$$\overline{y_a} = \frac{1}{m}\sum_{j=n+1}^{n+m} y_j,$$

$$\overline{y_b} = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

Before beginning the derivation of the Bayes factor, we note that the maximum likelihood estimates under M_1 are

$\alpha_{MLE} = \overline{y_a}$,

$\beta_{MLE} = \overline{y_b}$,

and the sum of squares of the residuals using the MLEs is

$SSR_{MLE} = (n+m)\overline{y^2} - m(\overline{y_a})^2 - n(\overline{y_b})^2$.

Prior distributions

For both models, we set the prior for (β, σ^2) to be

$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$.

For the extra parameter in M_1, we use the unit information prior centered at $\alpha = \beta$,

$p(\alpha|\beta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\alpha-\beta)^2\right]$.

Null model prior predictive distribution

The prior predictive probability of the data under M_0 is

$$p(y|M_0) = \int_0^\infty p(\beta, \sigma^2) \int_{-\infty}^{\infty} \prod_{j=1}^{n+n'} p(y_j|\beta, \sigma^2)\, d\beta\, d\sigma^2,$$

$$p(y|M_0) = (2\pi)^{-\frac{n+n'}{2}} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n+n'}{2}+1} \exp\left(-\frac{(n+n')\left(\overline{y^2} - (\bar{y})^2\right)}{2\sigma^2}\right)\left(\int_{-\infty}^{\infty} \exp\left(-\frac{n(\beta-\bar{y})^2}{2\sigma^2}\right) d\beta\right) d\sigma^2,$$

$$p(y|M_0) = (2\pi)^{-\frac{n+n'-1}{2}} (n+n')^{-\frac{1}{2}} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n+n'-1}{2}+1} \exp\left(-\frac{(n+n')\left(\overline{y^2} - (\bar{y})^2\right)}{2\sigma^2}\right) d\sigma^2,$$

$$p(y|M_0) = (2\pi)^{-\frac{n+n'-1}{2}} (n+n')^{-\frac{1}{2}} \Gamma\left(\frac{n+n'-1}{2}\right)\left[(n+n')\left(\overline{y^2} - (\bar{y})^2\right)\right]^{-(n+n'-1)/2}.$$

Alternative model prior predictive distribution

We define in advance:

$$E(\alpha|\beta, y) = \frac{(n'\overline{y_a} + \beta)}{n'+1},$$

$$\hat{\beta} = \left(\frac{(n+nn')\overline{y_b} + n'\overline{y_a}}{n+n'+nn'}\right),$$

$$SSR_1 = SSR_{MLE} + \frac{nn'}{(n+n'+nn')}(\overline{y_a} - \overline{y_b})^2.$$

As before, the effective sum of squares of the residuals under M_1 is the sum of the *SSR* using the maximum likelihood estimators and a penalty term for disagreement between the MLEs and the prior distribution.

Before dealing with the marginal probability of the data under M_1, we re-arrange the quadratic expression in α and β to ease the integrations.

$$(\alpha-\beta)^2 + \sum_{j=1}^{n}(y_j-\beta)^2 + \sum_{j=n+1}^{n+n'}(y_j-\alpha)^2$$

$$= \alpha^2 + \beta^2 - 2\alpha\beta + n\beta^2 - 2n\overline{y_b}\beta + n'\alpha^2 - 2n'\overline{y_a}\alpha + (n+n')\overline{y^2}$$

$$= (n+1)\beta^2 - 2n\overline{y_b}\beta + (n'+1)\alpha^2 - 2(n'\overline{y_a} + \beta)\alpha + (n+n')\overline{y^2}$$

$$= (n+1)\beta^2 - 2n\overline{y_b}\beta + (n'+1)\big(\alpha - E(\alpha|\beta,y)\big)^2 + (n+n')\overline{y^2} - \frac{(n'\overline{y_a}+\beta)^2}{n'+1}$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\beta^2 - 2\left(n\overline{y_b} + \frac{n'}{n'+1}\overline{y_a}\right)\beta + (n'+1)\big(\alpha - E(\alpha|\beta,y)\big)^2 + (n+n')\overline{y^2}$$
$$- \frac{n'^2}{n'+1}(\overline{y_a})^2$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\left[\beta^2 - 2\left(\frac{(n+nn')\overline{y_b} + n'\overline{y_a}}{n+n'+nn'}\right)\beta\right] + (n'+1)\big(\alpha - E(\alpha|\beta)\big)^2 + (n+n')\overline{y^2}$$
$$- \frac{n'^2}{n'+1}(\overline{y_a})^2$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\big(\beta-\hat{\beta}\big)^2 + (n'+1)\big(\alpha - E(\alpha|\beta,y)\big)^2 + (n+n')\overline{y^2} - \frac{n'^2}{n'+1}(\overline{y_a})^2$$
$$- \frac{[(n+nn')\overline{y_b} + n'\overline{y_a}]^2}{(n+n'+nn')(n'+1)}$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\big(\beta-\hat{\beta}\big)^2 + (n'+1)\big(\alpha - E(\alpha|\beta,y)\big)^2 + (n+n')\overline{y^2} - \frac{n'^2}{n'+1}(\overline{y_a})^2$$
$$- \frac{(n+nn')^2(\overline{y_b})^2 + n'^2(\overline{y_a})^2 + 2n'(n+nn')\overline{y_a}\,\overline{y_b}}{(n+n'+nn')(n'+1)}$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\big(\beta-\hat{\beta}\big)^2 + (n'+1)\big(\alpha - E(\alpha|\beta,y)\big)^2 + (n+n')\overline{y^2}$$
$$- \frac{n^2(n'+1)(\overline{y_b})^2 + n'^2(n+1)(\overline{y_a})^2 - 2nn'\overline{y_a}\,\overline{y_b}}{(n+n'+nn')}$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\left(\beta-\hat{\beta}\right)^2 + (n'+1)\left(\alpha - E(\alpha|\beta,y)\right)^2 + (n+n')\overline{y^2} - n'(\overline{y_a})^2 - n(\overline{y_b})^2 + \frac{nn'(\overline{y_b})^2 + nn'(\overline{y_a})^2 - 2nn'\overline{y_a}\,\overline{y_b}}{(n+n'+nn')}$$

$$= \left(\frac{n+n'+nn'}{m+1}\right)\left(\beta-\hat{\beta}\right)^2 + (n'+1)\left(\alpha - E(\alpha|\beta,y)\right)^2 + SSR_{MLE} + \frac{nn'}{(n+n'+nn')}(\overline{y_a} - \overline{y_b})^2$$

$$= \left(\frac{n+n'+nn'}{n'+1}\right)\left(\beta-\hat{\beta}\right)^2 + (n'+1)\left(\alpha - E(\alpha|\beta,y)\right)^2 + SSR_1.$$

The marginal probability of the data under M_1 is

$$p(y|M_1) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty p(\hat{a},\sigma^2)p(\alpha|\beta,\sigma^2)\prod_{i=j}^{n} p\left(y_j|\mu=\beta,\sigma^2\right)\prod_{i=n+1}^{n+n'} p\left(y_j|\mu=\alpha,\sigma^2\right)d\alpha\,d\beta\,d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-\frac{n+n'+1}{2}}\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{(n+n'+1)}{2}+1}\int_{-\infty}^\infty\int_{-\infty}^\infty \exp\left[-\frac{1}{2\sigma^2}\left((\alpha-\beta)^2 + \sum_{j=1}^{n}\left(y_j-\beta\right)^2 + \sum_{j=n+1}^{n+n'}\left(y_j-\alpha\right)^2\right)\right]d\alpha\,d\beta\,d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-\frac{n+n'+1}{2}}\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{(n+n'+1)}{2}+1}\exp\left(-\frac{SSR_1}{2\sigma^2}\right)\left(\int_{-\infty}^\infty \exp\left[-\frac{n+n'+nn'}{2\sigma^2(m+1)}\right]d\beta\right) \times\left(\int_{-\infty}^\infty \exp\left[-\frac{(m+1)}{2\sigma^2}\left(\alpha-E(\alpha|\beta,y)\right)^2\right]d\alpha\right)d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-\frac{n+n'-1}{2}}(n+n'+nn')^{-\frac{1}{2}}\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{(n+n'-1)}{2}+1}\exp\left(-\frac{SSR_1}{2\sigma^2}\right)d\sigma^2,$$

$$p(y|M_1) = (2\pi)^{-\frac{n+m-1}{2}}(n+n'+nn')^{-\frac{1}{2}}\Gamma\left(\frac{n+n'-1}{2}\right)(SSR_1)^{-(n+n'-1)/2}.$$

The Bayes factor is

$$BF = \frac{p(y|M_0)}{p(y|M_1)} = \sqrt{\frac{n+n'+nn'}{n+n'}}\left(\frac{SSR_1}{(n+n')\left(\overline{y^2}-(\bar{y})^2\right)}\right)^{\frac{n+n'-1}{2}}. \qquad \text{(S5)}$$

Equations (S3), (S4) and (S5) together are equivalent to equations (10), (11), and (12) of the main paper.